

RESEARCH

Open Access



Development of a machine-learning-based method for early fault detection in photovoltaic systems

Stylianos Voutsinas^{1*} , Dimitrios Karolidis¹ , Ioannis Voyiatzis¹  and Maria Samarakou¹ 

*Correspondence:
steliosvo@uniwa.gr

¹ Department of Informatics
and Computer Engineering,
University of West Attica, Athens,
Greece

Abstract

In the process of the decarbonization of energy production, the use of photovoltaic systems (PVS) is an increasing trend. In order to optimize the power generation, the fault detection and identification in PVS is significant. The purpose of this work is the study and implementation of such an algorithm, for the detection as many as faults arising on the DC side of a photovoltaic system. A machine learning technique was chosen. The dataset used to train the algorithm was based on a year's worth of irradiance and temperature data, as well as data from the PV cell used. The method uses logistic regression with cross validation as a new approach to detect and identify faults in PVS. It is applied to smart PV arrays, that can transmit voltage and current measurements from each PV cell of the array individually. The results are satisfactory since the algorithm can detect the majority of faults that occur on the DC side of a photovoltaic (open-circuit fault, short-circuit fault, mismatch faults). The accuracy of the algorithm (97.11%) is comparable to other methods presented by the literature. Moreover, the computational cost of the proposed method is significantly lower than the methods presented in the literature. In summary, the performance of the implemented algorithm is considered particularly satisfactory and can be easily applied to PVS.

Keywords: Photovoltaic systems, Photovoltaic fault detection algorithms, I–V curves, Machine learning

Introduction

According to the report of the International Energy Agency (IEA) for the year 2021, approximately 81% of global electricity production is based on the combustion of coal, oil, and natural gas. Within a year, the use of alternative energy sources such as the use of photovoltaics and wind turbines [1] has increased by 1% [2]. In the European Union, energy production from PVS during the years 2008–2020 increased by 1848% [3]. This increase can be explained due to the ability of the PVS to zero carbon footprint—therefore their use is in line with the Paris Agreement. Furthermore, PVS are easy to install [4–7]. However, it should be noted that their low efficiency and low-profit margin per MWh are deterrents for large investments in PVS [8, 9]. With the progress of embedded systems, the transition to smart photovoltaic systems is gradually taking place. Smart

PVS through power line communication (PLC) can maximize the energy production of a PVS, providing additional control and parameterization of both the array itself, but also fault control at the PV cell level [10].

The advantages of machine learning (ML) methods over other artificial intelligence (AI) and threshold-based methods are many and include their data-driven nature, scalability, automation, continuous learning, and predictive accuracy [11, 12]. ML algorithms are designed to learn from data and make predictions based on patterns in the data, rather than relying on pre-programmed rules. This feature allows to the ML-based algorithms for more accurate predictions and decision-making [13]. Unlike other AI methods, which can be limited by pre-programmed rules, ML algorithms can handle large amounts of data and are suitable for processing big data sets, making them scalable [14]. However, it should be noted that several AI algorithms such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), gradient boosting machines (GBMs), and rule-based systems can be scaled up to handle large datasets and complex problems. Automation is another advantage of ML algorithms as they can automate many tasks that would otherwise require human intervention, leading to increased efficiency and reduced costs [12]. In addition, ML algorithms can continue to learn and improve over time with new data, making them more adaptive and versatile than traditional AI methods [11].

Compared to threshold-based methods, ML algorithms can make more complex decisions based on patterns in the data, leading to improved predictive accuracy. Threshold-based methods rely on pre-defined thresholds to make decisions, which can lead to oversimplification and decreased accuracy [13, 14]. Unlike threshold-based methods, ML algorithms can also handle non-linear relationships in the data, making them more suitable for a wider range of applications [12].

In conclusion, the advantages of ML methods over other AI approaches and threshold-based methods make them a powerful tool for prediction and decision making in many fields. However, the specific advantages of ML will vary depending on the application. Fast execution time and low memory usage are crucial for the success of machine learning (ML) algorithms in real-world applications [12, 15]. ML algorithms can be computationally intensive. Slow execution times can lead to increased processing times and decreased efficiency [15]. In addition, many real-world applications require the use of large amounts of data, and the memory requirements of ML algorithms can be substantial [12]. If the memory requirements of the algorithm are too high, it may not be feasible to run the algorithm on the available hardware, leading to decreased performance and accuracy [15]. Therefore, fast execution time and low memory usage are essential for the successful deployment of ML algorithms in real-world applications, as they ensure that the algorithms are computationally feasible and efficient [12]. To sum up, the objectives of this work is to implement a machine learning-based fault identification and detection algorithm, capable.

- a) To detect at least the three main categories of faults (open-circuit fault, short-circuit fault, mismatch faults) that arise on the DC side of a PVS,
- b) Of small computational cost. It should have small execution time per prediction, while in parallel, it should consume minimum memory.

- c) Of high accuracy.
- d) To be applied to smart PV arrays which can transmit voltage and current measurements from each PV cell of the array individually. In this way, the operation of each cell of the smart PV array is monitored [10, 16]. When there is a flaw, the faulty PV cell is isolated from the PV array.

The structure of this paper is as follows: Sect. 2, includes the similar works presented in the literature. A summary of the most common types of faults that occur on the DC side of a PVS will be presented. Following that, techniques for fault detection proposed in the literature will be assessed for their accuracy, memory and time requirements, and finally, their ability to detect as many unique types of faults as feasible. Methods that can detect and identify a wide variety of defects will be preferred above those that do not meet the selected criterion. In Sect. 3 the methodology of the developed method is presented. In Sect. 4 the results from the experimental procedure and the discussion of this paper are presented. The findings of the suggested approach will also be discussed and compared to other methods proposed in the literature. Finally, in Sect. 5 the conclusion of this research is presented.

Literature review

Faults in PVS

The main purpose of fault detection and classification methods is to identify what is causing fluctuations in the energy production of a PVS [17]. Different types of faults can occur on both the AC and DC sides of a PVS [18]. Traditional protection systems are designed to address AC faults, but faults on the DC side can be harder to identify and fix [17, 19]. Typical faults on the DC side of the PVS are shown in Fig. 1 and briefly presented in Table 1.

One common category of DC faults is the mismatch faults, which can significantly reduce the power output of a PVS. Mismatch faults can be temporary or permanent. Temporary mismatch faults can be caused by particle accumulation on the surface of a PVS such as dust, bird droppings, or from the shading of the PVS due to some tree or some cloud. Permanent mismatch faults can be caused by damage to the adhesive materials, surface cracks on the PVS, gaps between layers of the PV module that cause shading, or deterioration of the semiconductor material [21]. It should be noted that permanent mismatch faults can occur in a system even as a result of another fault, such as an open circuit fault. Short-circuit faults can also occur when there are problems with the connections in a PVS, leading to the unintended connection between two points of the PVS [22]. An unintended short circuit between two voltage potentials across two neighboring strings or between two voltage potentials inside a single string [23], is called line-to-line fault. If the short-circuit involves the connection of a current-carrier with a non-current carrier, such as the PV frame then the fault is named ground fault or line-to-ground fault [24].

Open-circuit faults can occur when there is a disconnection on a PV string [25] (usually caused by poor soldering), but under certain conditions an open-circuit fault can also lead to arc failures, leading to high-frequency noise and rapid decreases in output voltage and current [26]. It is worth noting that arc faults can be mitigated using an Arc

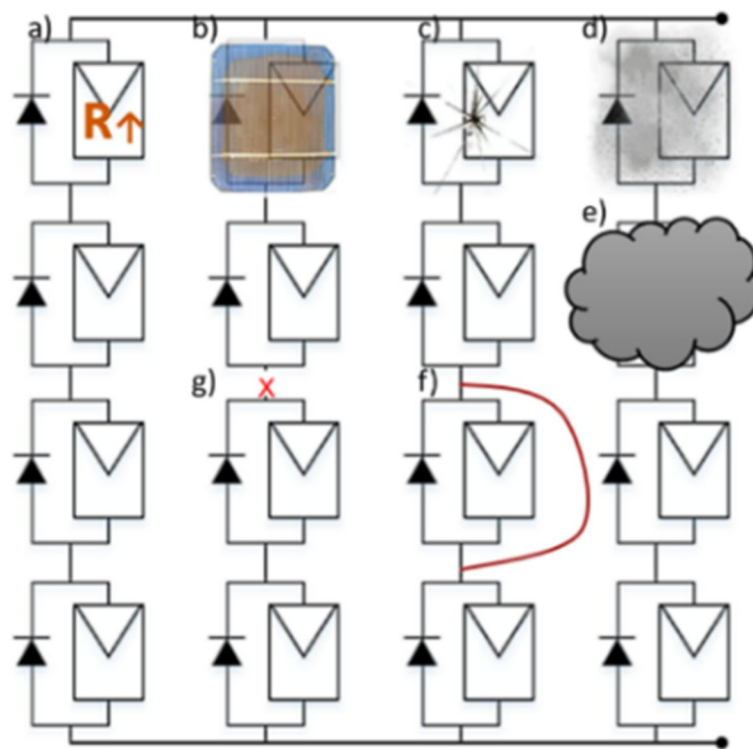


Fig. 1 Manifestation of common types of faults on the DC side of a PVS. **a** Degradation of the semiconductor. **b** Discolorations. **c** Microcracks. **d** Particles accumulation. **e** Shading. **f** Short-Circuit. **g** Open-circuit

Table 1 Common types of dc faults and their causes [20]

Type of fault	Manifestation	Cause
Permanent mismatch	Degradation of modules Glass breakage Interconnect breakage of busbars Defects in frame Cell breakage Microcracks	Low-quality materials Human errors Poor handling Side-product of another fault
Temporary mismatch	Partial shading Snow covering Dust/bird dropping/leaves	Environmental
Arc	Series/parallel arc fault	Corrosion Insulation damage Open-circuits
Line to line	Between two v. potentials on different PV strings, or in the same PV string	Unintentional connections due to short-circuit
Short circuit	Short circuit between two points	Bad connections, defects
Open circuit	open circuit between two points (mostly on blocking diode (BcD) and bypass diode (BpD))	Loose connections
Ground	Short-circuit to the ground	Non-current-carrying conductors (such as PV racks and frames that are grounded) are shorted with current-carrying conductors

Fault Circuit Interrupter (AFCI) and ground faults can be monitored using a Residual Current Monitor (RCM) [27–29]. Special mention is made of both arc faults and ground faults, because both are particularly dangerous. The former can cause a fire, the spread

of which can threaten the entire installation, while the latter can turn the PVS frames into live traps for the installation's personnel, putting their lives at risk [18, 30]. Using high-quality materials and proper handling during the transport and installation of a PVS can also help reduce the risk of mismatch faults [31] since a proper installation avoids microcracks on the PVS surface and the use of better quality materials will greatly slow down the appearance of discoloration.

Fault detection algorithms

There are various techniques in the literature in order to detect faults in PVS. They can be categorized into three main groups; electrical characterization, visual inspection and thermal imaging. Visual inspection techniques [32, 33] require regular inspections to detect anomalies in the appearance of the PVS, thus they cannot be used for real-time monitoring. Thermal imaging techniques [34–37] involve the use of specialized equipment that increases the cost of PVS installation. Electrical diagnostic methods, on the other hand, can be performed either on-site or remotely. They are based on monitoring the specific electronic signatures that each fault produces and its effect on the output power of the PVS [38]. Many electrical diagnostic methods base their operation data analysis on the I-V curve of the PVS, which can detect various faults [39].

Several machine learning (ML)-based techniques with high fault detection accuracy have been published in the literature. Many of these methods are trained using I-V curve data. Although ML-based approaches necessitate a significant amount of processing power for training, their capacities to self-learn and adapt to a variety of inputs overcome the drawback of the required processing power [40].

The method presented by Chen et al. [41] can identify a wide range of faults. It is based on a kernel-based extreme learning machine (KELM), which has seven inputs (V_{oc} , I_{sc} , V_{mpp} , I_{mpp} , α_1 , R_s , and RMSE values) and five outputs (4 faulty states and a normal state). Harrou et al. [42], in order to develop the single diode model for the monitored PV cells, employed an artificial bee colony (ABC) method to handle irradiance and temperature data. At the maximum power point, the current, voltage, and power levels are determined. The disparity between simulated and measured values is utilized to indicate the presence of a fault. Voutsinas et al. [43], used data from I-V curves to train a multi-output feed-forward neural network consisting of $8 \times 10 \times 10 \times 6$ neurons. This implementation has 4 faulty states and a normal state. To identify line-to-line faults, Yi and Etemadi [44] used a multi-resolution signal decomposition (MSD) combined with a support vector machine (SVM) for feature extraction. Xia et al. [45] used wavelet decomposition in conjunction with SVM to identify series DC arc defects. Harrou et al. [46] used a binary SVM classifier to detect irregularities in output DC and power using a PSIM simulation of an installed grid-connected PVS. Wang et al. [47] employed a multi-class SVM to identify and categorize line-to-line faults and anomalous degradation faults in a PV module. Winston et al. [48] utilized a feed forward back propagation neural network combined with an SVM in order to detect micro-cracks and hotspots. Yi and Etemadi [49] proposed a method for detecting line-to-line and line-to-ground faults, which was based primarily on the use of a multi-resolution signal decomposition (MSD) algorithm on a fuzzy inference system. Memon et al. [50] proposed the use of a convolutional neural network (CNN) that used parameters such as irradiance temperature voltage and

current, in order to detect the presence of faults. Jia et al. [51], presented a near perfect accuracy method for detecting arc faults using logistic regression. Fadhel et al. [52] proposed a data driven approach for detecting faults caused by shading on a PVS. The method is based in principal component analysis (PCA) that used data from I-V curves to detect faults with significant accuracy. Finally, Dai et al. [53] suggested a deep reinforcement learning-based PVS fault detection technique. The starting premise for this approach is data-driven. The fault diagnostic model of the PVS is created, and the deep neural network is used to estimate the decision network in order to find the optimum strategy, allowing the photovoltaic power generation system to be fault diagnosed.

The need to improve the reliability and performance of a smart PV array is the motivation for the development of a rapid and accurate fault detection and identification method based on ML. In the context of renewable energy systems, fault detection and identification are crucial for ensuring optimal energy generation and preventing catastrophic failures. However, traditional fault detection methods are often time-consuming and rely on human intervention, which can lead to delayed or inaccurate diagnoses. By leveraging the power of machine learning, the proposed approach can quickly and accurately identify faulty PV cells in real-time, based on the data transmitted by each cell. While there are limitations to machine learning, such as overfitting, the use of a rigorous cross-validation process can help mitigate these issues and improve the accuracy of the model. Therefore, the use of machine learning in fault detection and identification for smart PV arrays is a promising approach that can improve system reliability and performance.

Methods

To create the dataset, irradiance and temperature data are required as well as a model that will simulate the operation of each photovoltaic cell. The electrical output of a photovoltaic cell can be approximated by an analogous model circuit named single-diode model (SDM) with five parameters; these parameters are unknown and required to predict the performance of the PV module and are derived from the photovoltaic cell's current equation for a given temperature and irradiance. Both models can simulate PV cell performance in low voltage and/or high external temperature circumstances [54]. Equation (1) denotes the current equation for the single-diode model, whereas Fig. 2 depicts the analogous circuit. In Eq. (1), I_{ph} is the current generated by the irradiance of light, I_{o1} is the reverse saturation current of diode D_1 , q is the electron charge, k is the Boltzmann's constant, α_1 is the diode's ideality factor, T is the temperature expressed in Kelvin degrees and R_s , R_{sh} are the resistors in series and shunt. The approach provided in the

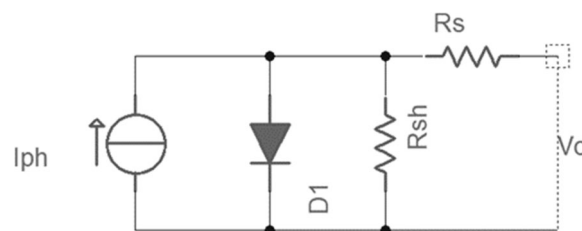


Fig. 2 Single-diode model, equivalent circuit

work of De Soto et al. [55] is used to determine the five parameters of Eq. 1 (I_{ph} , I_s , α_1 , R_s , and R_{sh})

$$I = I_{ph} - I_0 \left(e^{\frac{q(V+I R_s)}{a_1 k T}} - 1 \right) - \frac{V + I R_s}{R_{SH}} \tag{1}$$

Looking at Fig. 3, plots depicted in green represent the I_{mpp} and V_{mpp} values according to the manufacturer’s datasheet, while plots in yellow represent the values created by the application that will create the dataset, based on De Soto method. It should be noted that the deviations in the voltage are less than 10 mV and the corresponding ones in the current are less than 100 mA, which makes them negligible.

Since the SDM is fully functional, the next step is to collect irradiance and temperature data. The PVGIS service [56] is used to acquire irradiance and temperature data

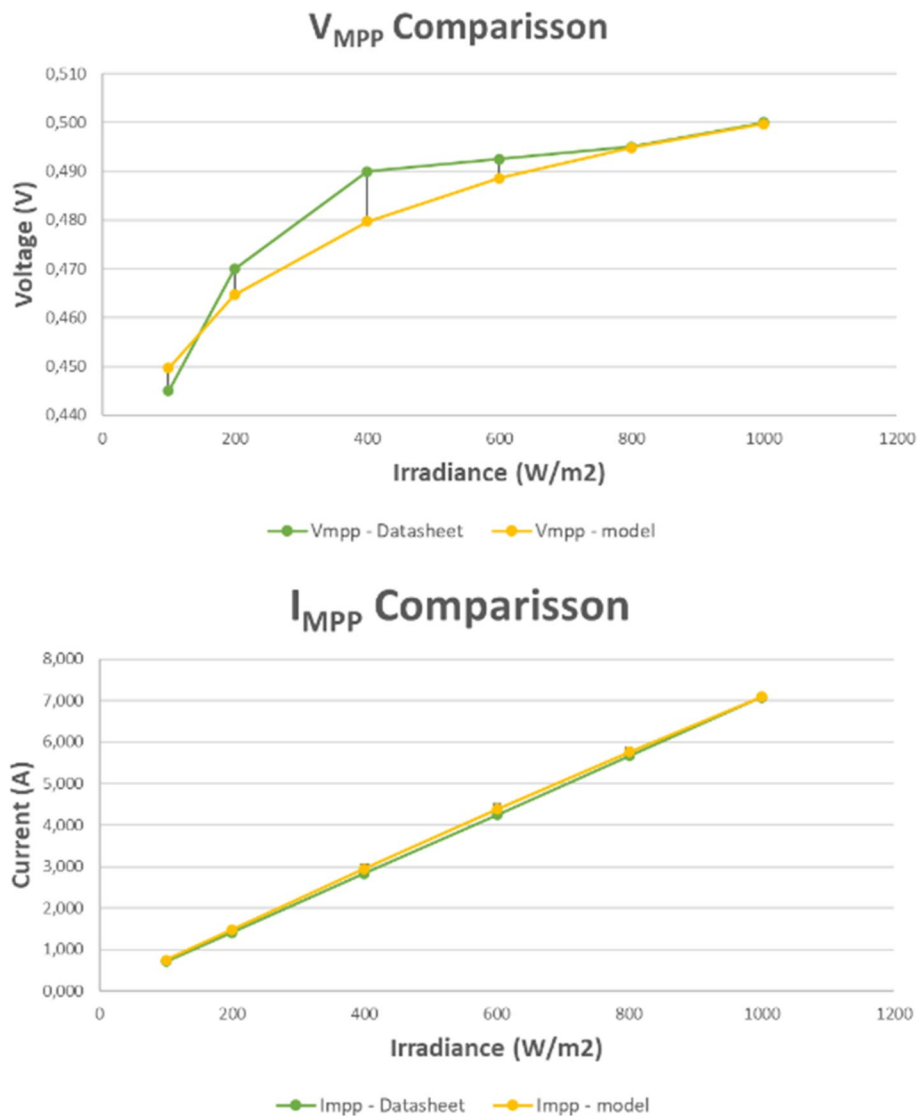


Fig. 3 Comparison between the values of I_{mpp} and V_{mpp} from the datasheet of the PV cell and the I_{mpp} and V_{mpp} generated by the SDM

Table 2 Operational status code encodings–conditions

Op. status code	0	1	2	3	4
<i>Fault type</i>	Normal operation	Short-circuit	Open-circuit	Mismatch	Undefined
<i>Status condition</i>	$P_{cell} \geq 0.9 P_{mpp-model}$	$I_{cell} \geq 0.9 I_{scmodel}$ and $V_{cell} \leq 0.1 V_{ocmodel}$	$I_{cell} \geq 0.1 I_{scmodel}$ and $V_{cell} \leq 0.9 V_{ocmodel}$	$FF_{model} - FF_{cell} \geq 0.1$	If the preceding status conditions do not apply

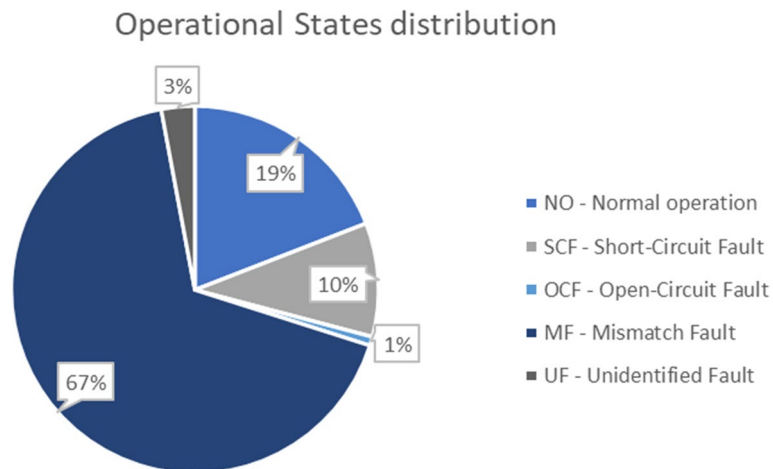


Fig. 4 Distribution of operational states within the dataset

for the Egaleo region (Attica, Greece). Twelve files were provided from PVGIS, 1 for each month, from January till December, indicating the average hourly temperature and irradiance for all days of each month. These files were consolidated and the corresponding records to the certain hours after sunset and before sunrise were discarded. The concatenated yearly file served as input to a Python script, which allowed the SDM to construct I–V curves for the 147 pairs of irradiance and temperature. From the 147 pairs, temperature varies from 1.32 to 35.06 °C, while irradiance varies from 0.04 to 984.84 W/m².

After all the data have been collected, the last script undertakes the interconnection of the data and the creation of 110,250 records wide dataset. Each record has the format: Temperature (°K), GHI (W/m²), $V_{mppModel}$ (V), $I_{mppModel}$ (A), $V_{ocModel}$ (V), $I_{scModel}$ (A), V_{Cell} (V), I_{Cell} (A), and operational status. Temperature and GHI values are retrieved directly from the PVGIS service. $V_{mppModel}$, $I_{mppModel}$, $V_{ocModel}$, and $I_{scModel}$ are retrieved from each generated I–V curve. V_{Cell} and I_{Cell} are voltage and current values ranging from 0- $V_{ocModel}$ and 0- $I_{scModel}$, and are measurements made in each PV cell of the array.

The operational status codes are encoded according to Table 2. While Fig. 4 depicts the percentages of the operational status codes within the full dataset, and in Fig. 5 the nine features of the dataset are grouped by their operational status code.

Observing the values of the last field of the dataset (operational status), we are led to the conclusion that this is a multi-class classification problem. To turn a multi-class problem into a set of binary tasks, the use of either one-vs-one (OVO) or one-vs-rest

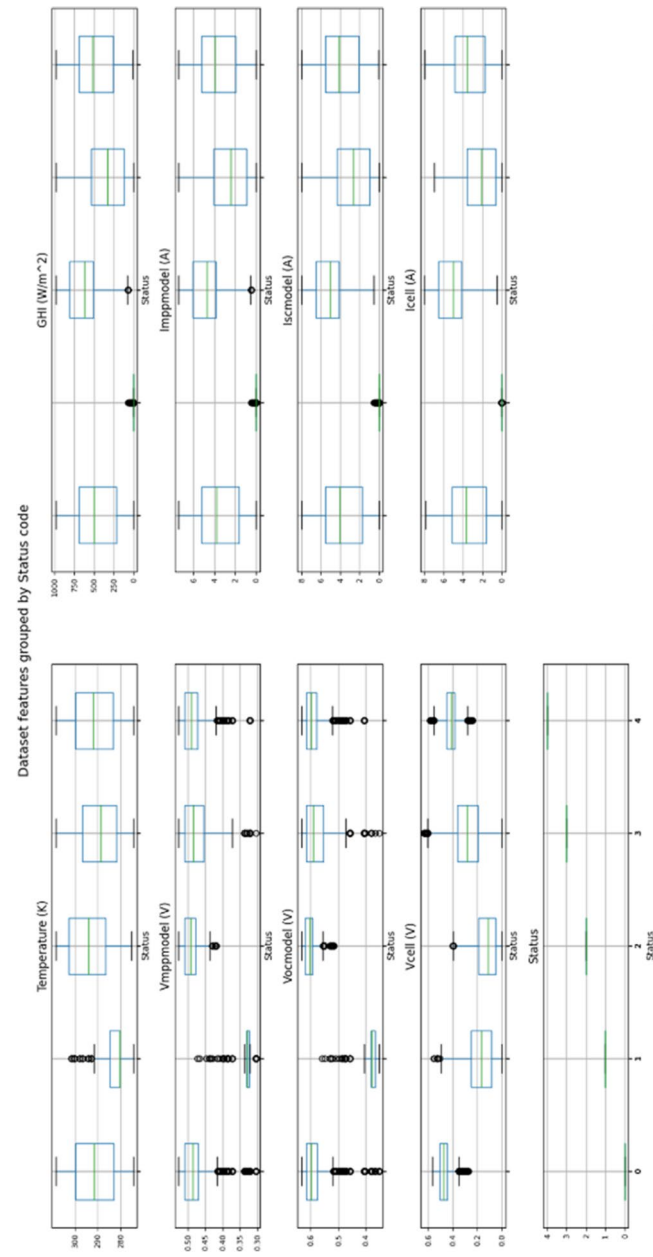


Fig. 5 Dataset features grouped by operational status code

(OVR) strategies is suggested. Using the OVO strategy requires 10 classifiers ($[N*(N-1)]/2$), while for the OVR strategy only 5 are required. Each operational status in the dataset has its classifier. Consequently, the second strategy is considered preferable.

The use of logistic regression enhanced with CV for fault detection in photovoltaic systems can be considered significantly constructive. While logistic regression is a well-established machine learning technique for binary classification tasks [51], its use in the specific context of fault detection in PV systems is still developing. The addition of CV to the logistic regression model can be considered innovative as it helps to improve the performance of the model by reducing overfitting and increasing its ability to generalize to new data. By incorporating CV into the logistic regression model, the effectiveness of the model for fault detection in PV systems is likely to be improved, which could lead to new and more effective approaches for identifying faults in PV systems. LogisticRegressionCV classifier is compatible with the OVR strategy. It is similar to plain logistic regression but it has been hyperparameter-tuned (through CV). It tries several regularization strengths and chooses the optimal one based on CV ratings then refits a single model on the entire training set, using that best C (Inverse of regularization strength). The LogisticRegressionCV parameters were determined as follows: The trainer continues training for 10,000 iterations to find better weights. The number of CV sets is set to 3. The 'ovr' option has been selected, to follow the one-vs-rest strategy. Since it is a multi-class problem, the limited-memory BFGS (LBFGS) optimization algorithm was chosen as a solver combined with the regularization parameter (penalty) ridge regression (L2). Finally, the size of the list of the available values (C_s) for the coefficient of the inverse of regularization strength (C) is set to 10.

The development of the application as well as the creation of the dataset was done in Python 3.9 language using sklearn 1.1.2 [57] and PVlib 0.9 libraries [58]. The photovoltaic cell used in the dataset is the Solar Cells Hellas SCH6P-60 Multicrystalline Solar Cell [59].

Results and discussion

In previous sections, the development of a machine-learning algorithm based on logistic regression with cross-validation, capable of detecting and identifying faults in the DC side of a PVS was presented. Then the experimental measurements of the method are listed and discussed. The methods presented in the literature review are compared with each-other and with the method presented.

The experimental process was performed on an AMD Ryzen 3 5400U processor, 8.00 GB DDR4 RAM and PCIe M.2 SSD. During the measurement process, no other processes were running in the foreground of the operating system apart from the basic processes of the OS. This was done in order the results to be as accurate as possible.

The experimental data have been divided into two tables (Tables 3 and 4), Table 3 has the qualitative characteristics and Table 4 has the quantitative characteristics.

Regarding the quantitative characteristics of the measurements. The average training time and memory required for the training process is shown in Table 4. In more detail

Table 3 Experimental results—qualitative characteristics

Accuracy	Precision	AUC	Recall	F1-score
97.11%	0.955	0.998	0.945	0.949

Table 4 Experimental results—quantitative characteristics

Training time	Memory required	Execution time/prediction call	Memory/prediction call
318 s	9.1 MB	180 KB	8 ms

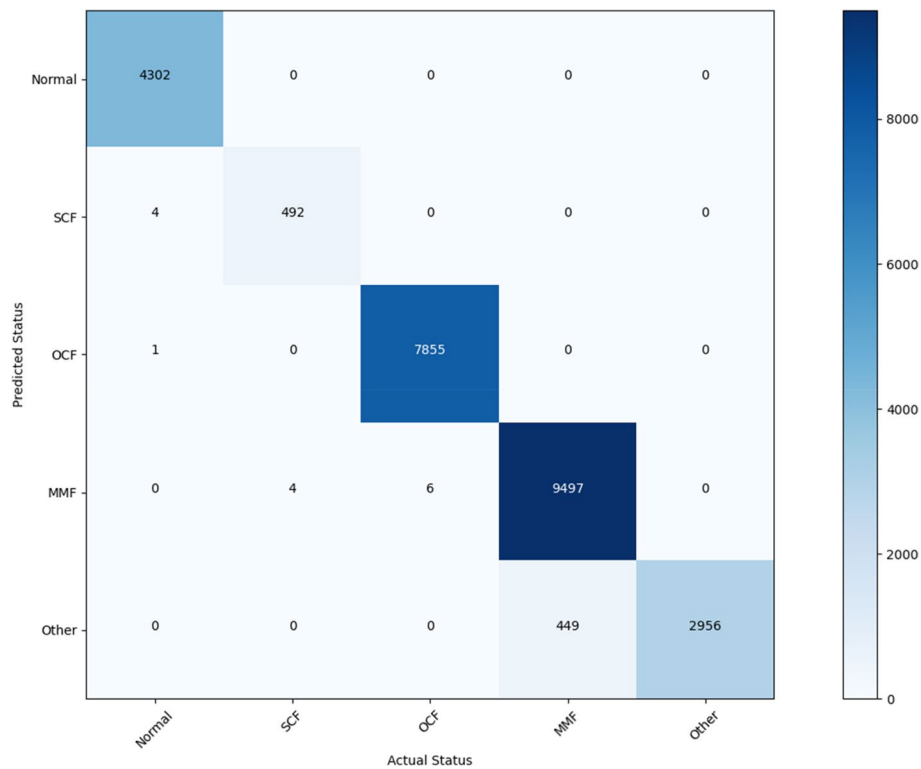


Fig. 6 5 × 5 confusion matrix

the logistic regression took 113 s and consumed 6.68 MB of memory. The process of fitting and training the model lasted 205 s and consumed 2.48 MB of memory. Accordingly, the use of the model in order to make a forecast of the operating state of the PV cell is 8 ms/prediction call with a memory consumption of 180 KB/prediction call. It should be noted that in order to obtain the measurements concerning the execution time, but also the memory that we reserved when calling the generated model of the method, a loop of 100 iterations was used, and the above values (execution time, memory consumption) are essentially the average values of 100 iterations.

In terms of the measurements’ qualitative features, Fig. 6 shows the confusion matrix after the classification and the fitting process. Here, we can observe the true positive predictions, the true negative predictions as well as type 1 and 2 errors (false positive, false negative). While Fig. 7 depicts the receiver operating characteristic (ROC) curves for the five classifier that were used. These curves display the performance of each classifier across all categorization criteria. The greater the area under the curve (AUC), the better the model distinguishes across classes. We have an AUC of 99.8% based on the experimental data. This is almost an ideal circumstance. The data from true positives (TP) and true negatives (TN) overlap by less than 0.2%.

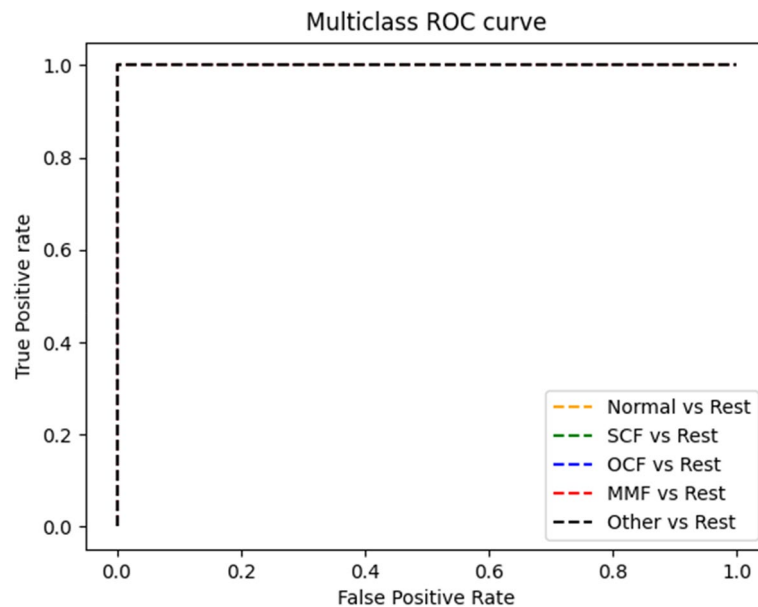


Fig. 7 ROC curves for the five classifiers

When TP and TN do not overlap, the model provides an ideal measure of separability. This means that each one of the classifiers used, nearly precisely, differentiates between its positive and negative classes.

The F1-score (0.949) is an improved version of two simpler performance metrics: accuracy and recall. Precision (0.955) that indicates the proportion of anticipated positives that are genuinely positive, while recall (0.945) indicates the proportion of actual positives that were accurately detected. It is commonly referred to as the harmonic mean of the two metrics. The goal is to produce a single metric that evenly weights the two values (precision and recall). The accuracy statistic (97.11%) indicates how many times the model predicted correctly over the full dataset.

The measurements presented in Table 4 are all remarkably high, a fact that makes the algorithm particularly reliable.

Table 5 below shows the comparison of the implemented method compared to the methods presented in the literature. The comparison among the methods is based on the accuracy of each method, the ability to identify the three main categories of faults on the DC side of a PVS, but also on the computational cost of memory usage and the execution time using each method. In Table 5 additionally, in the field of “comments”, other characteristics, advantages, or disadvantages of each method are presented.

According to Table 5, the developed method will be compared with other thirteen methods that were presented in the last 5 years (2017–2022) in the literature. Six methods [41–43, 46, 50, 53], can detect the three main types of faults on the DC side of a PVS (open-circuit fault, short-circuit fault, mismatch faults). One method [47] can detect two types of faults (short-circuit fault, mismatch faults), and six methods [44, 45, 48, 49, 51, 52] can detect only one type of fault (open-circuit fault, arc fault, mismatch faults, short-circuit fault, arc fault, mismatch faults respectively). Regarding the accuracy of fault detection, nine methods [41, 42, 45, 47–51, 53] show an accuracy

Table 5 Comparison of the methods

Method/Ref num	OCF	SCF	MF	Other	Mem/ call (KB)	Ex. time/ call (ms)	Accuracy/comments
Developed method [-]	✓	✓	✓	✓	180	8	97.11%
Chen et al. [41]	✓	✓	✓	-	-	-	98.8%
Harrou et al. [42]	✓	✓	✓	-	-	-	$r^2 > 0.9$ RMSE 0.14 MAPE2.83%
Voutsinas et al. [43]	✓	✓	✓	✓	220	44	93.4%
Yi and Etemadi [44]	-	✓	-	-	-	-	94.74%
Xia et al. [45]	-	-	-	-	-	-	Arc fault detection 96%
Harrou et al. [46]	✓	✓	✓	-	-	-	Ranges from 89.6–98% depending on the fault it detects
Wang et al. [47]	-	✓	✓	-	-	-	97.78%
Winston et al. [48]	-	-	✓	-	-	-	99%
Yi and Etemadi [49]	-	✓	-	-	-	-	Ranges from 97.69 to 100% depending on the fault it detects
Memon et al. [50]	✓	✓	✓	-	-	160–70	95.2%. Execution time varies according to the kind of the detected fault
Jia et al. [51]	-	-	-	-	-	-	Arc fault detection 100%
Fadhel et al. [52]	-	-	✓	-	-	-	Different types of shading faults. Insensitive to sudden irradiance variations 88%
Dai et al. [53]	✓	✓	✓	-	-	-	96.6%

of more than 95%, while one method [46] shows fluctuating accuracy depending on the type of the fault (89.6–98%). Finally, only two methods [43, 50] provide information about their computational cost, it has to be noted though that the method presented in [50] provides data only for its execution time.

Comparing the new method implemented and presented in this work with the corresponding 13 methods from the literature, the method can perceive the three basic categories of faults that may occur, while in addition, it can also perceive the existence of other errors. The fourth category of faults that can be perceived by our method refers to faults that do not correspond to any of the three basic categories of errors that we study, but they can be considered as the transition from normal operation to some faulty state, thus their monitoring can act as a warning indicator of future problems. As far as the accuracy of its measurements is concerned, it has a performance greater than 95% (more specifically 97.11%) and provides information about its computational cost in memory and execution time (180 KB RAM per call, 8 ms per call). It should be noted that most of the methods presented in the literature do not report information about the execution time and the memory they consume. The best execution time is shown in the presented method (8 ms), with second best performance that presented in [43] (44 ms), and with third best performance that presented in [50] (160–70 ms; note: the execution time the operating status of the PVS changes accordingly). The execution time is objectively related to the hardware of the computer on which the method is executed, but methods

based on ANNs and CNNs, due to the complexity of their models, lead to an increase in the execution time. On the other hand, the method presented in the current work is based on logistic regression, a simple and fast machine learning method compared to other complex models like deep neural networks. This is because logistic regression is a linear model, which means that it has a relatively small number of parameters and can be trained relatively quickly. Logistic regression can be executed very quickly, even on large datasets, due to its linear nature and the efficient optimization algorithms that are available for training the model. Due to this, the presented method exhibits the shortest execution time.

The developed method is considered to have comparable response in relation to the corresponding methods in the literature, but in some criteria such as the identification of additional faults, or in the part of the computational cost it goes beyond the limits set by the literature. Its installation in a PVS with quality materials in order to limit the appearance of permanent mismatch faults, while at the same time having AFCI and RCM sub-units installed to the PVS in order to immediately detect arcing and current leakage to ground, will guarantee the smooth operation of the PVS.

As a future work, it is proposed to increase the records of the dataset, in order to further increase the accuracy of the method. Furthermore, the additional categorization of the faults with the use of appropriate hardware will be able to separate the subcategories of the faults, such as for example the separation between permanent and temporary mismatch faults.

Conclusions

The purpose of this work is to design an algorithm for the early detection and identification of faults that may occur in the DC part of a PVS. The results of the method are particularly encouraging since it can identify with 97.11% accuracy the three main categories of faults (open-circuit fault, short-circuit fault, permanent and temporary mismatch faults, and can detect the presence of extra undefined faults) on the DC side of a PVS. The latter fault category includes faults that do not belong to any of the three main categories of faults that appear on the DC side of a PVS, or they can signify the transition from the normal operation of the photovoltaic cell to some faulty state. Furthermore, comparing our method with other methods introduced in the literature, our method is quick and memory-efficient when used for output prediction (180 KB RAM per call, 8 ms per call). Comparing our method with the existing methods from the literature, it provides similar levels of accuracy, while in the majority of cases it identifies more faults. It should not be overlooked that the specific method can be applied to typical PVS installations (with minor modifications), not only to smart PVSs. In fact, in the latter, our method can be used, at the photovoltaic string level but also at the PV cell level, which is very important since it gives full real-time control over the state of each cell of the PVS. The results indicate that it can be used in PVS-based power plants.

Abbreviations

AFCI	Arc fault circuit interrupter
I_{cell}	Current measurement from PV cell
ANN	Artificial neural network
I_{mpp}	Maximum power point current
AUC	Area under curve

I_{ph}	Generated photo-current
CV	Cross-validation
I_s	Saturation current
DC	Direct current
I_{sc}	Short-circuit current
FF	Fill factor
K	Boltzmann's constant
IEA	International Energy Agency
P_{cell}	Power measured from PV cell
GHI	Global Horizontal Irradiance
P_{mpp}	Power in maximum power point
KELM	Kernel-based extreme learning machine
Q	Electron charge
M_{pp}	Maximum power point
R_s	Series resistance
MSD	Multi-resolution signal decomposition
R_{sh}	Shunt resistance
OVO	One vs one
OVR	One vs rest
T	Temperature expressed in Kelvin degrees
PCA	Principal component analysis
V_{cell}	Voltage measurement from PV cell
PV	Photovoltaic
V_{mpp}	Maximum power point voltage
PVS	Photovoltaic systems
V_{oc}	Open-circuit voltage
RCM	Residual current monitor
α_1	Diode ideality factor
RMSE	Root-mean-square error
ROC	Receiver operating characteristic
SDM	Single diode model
SVM	Support vector machine
TN	True negative
TP	True positive

Acknowledgements

Not applicable.

Authors' contributions

All authors have contributed equally to the work. The authors read and approved the final manuscript.

Funding

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation, under the call RESEARCH–CREATE–INNOVATE (project code: T1EDK-01485).

Availability of data and materials

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 15 February 2023 Accepted: 11 April 2023

Published online: 18 April 2023

References

1. Bharatee A, Ray PK, Ghosh A (2022) A power management scheme for grid-connected PV integrated with hybrid energy storage system. *J Mod Power Syst Clean Energy* 10(4):954–963. <https://doi.org/10.35833/MPCE.2021.000023>
2. International Energy Agency (IEA) (2021) "Key world energy statistics 2021 – statistics report". [Online]. Available: <https://iea.blob.core.windows.net/assets/52f66a88-0b63-4ad2-94a5-29d36e864b82/KeyWorldEnergyStatistics2021.pdf>
3. Eurostat, Wind and water provide most renewable electricity; solar is the fastest-growing energy source, 2022. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Renewable_energy_statistics#Wind_and_water_provide_most_renewable_electricity.3B_solar_is_the_fastest-growing_energy_source
4. UNFCCC (2015) The Paris Agreement – Publication

5. Peng J, Lu L, Yang H (2013) Review on life cycle assessment of energy payback and greenhouse gas emission of solar photovoltaic systems *Renew Sustain. Energy Rev* 19:255–274. <https://doi.org/10.1016/j.rser.2012.11.035>
6. Mundada AS, Prehoda EW, Pearce JM (2017) U.S. market for solar photovoltaic plug-and-play systems. *Renew Energy*. 103:255–264. <https://doi.org/10.1016/j.renene.2016.11.034>
7. Fouly SEA, Abidin AR (2022) A methodology towards delivery of net zero carbon building in hot arid climate with reference to low residential buildings — the western desert in Egypt. *J Eng Appl Sci* 69(1):46. <https://doi.org/10.1186/s44147-022-00084-6>
8. NREL (2022) Best Research-Cell Efficiencies. Who, USA
9. Lugo-laguna D, Arcos-Vargas A, Nuñez-hernandez F (2021) A european assessment of the solar energy cost: key factors and optimal technology. *Sustain* 13(6):1–25. <https://doi.org/10.3390/su13063238>
10. S. Voutsinas, I. Mandourarakis, E. Koutroulis, D. Karolidis, I. Voyiatzis, and M. Samarakou, Control and communication for smart photovoltaic arrays, in 26th Pan-Hellenic Conference on Informatics (PCI 2022) 2022;6, <https://doi.org/10.1145/3575879.3575983>
11. Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning*. Springer, New York
12. I. Goodfellow, Y. Bengio, and A. Courville (2016) *Deep Learning*. The MIT Press, ISBN: 9780262035613
13. Alpaydin E (2014) *Introduction to Machine Learning*, 3rd edn. MIT Press, Cambridge, MA
14. C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics), 1st ed. Springer, 2007.
15. Bottou L (2012) *Stochastic gradient descent tricks*. Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp 421–436
16. S. Voutsinas, D. Karolidis, I. Voyiatzis, and M. Samarakou, Development of an IoT power management system for photovoltaic power plants, in 2022 11th International Conference on Modern Circuits and Systems Technologies (MOCASST), 2022; 1–5, <https://doi.org/10.1109/MOCASST54814.2022.9837652>.
17. Monadi M, Zamani MA, Candela JI, Luna A, Rodriguez P (2015) Protection of AC and DC distribution systems Embedding distributed energy resources: a comparative review and analysis. *Renew Sustain Energy Rev* 51:1578–1593. <https://doi.org/10.1016/j.rser.2015.07.013>
18. Hong Y-Y, Pula RA (2022) Methods of photovoltaic fault detection and classification: A review. *Energy Rep* 8:5898–5929. <https://doi.org/10.1016/j.egyr.2022.04.043>
19. Huang JM, Wai RJ, Gao W (2019) Newly-designed fault diagnostic method for solar photovoltaic generation system based on IV-Curve measurement. *IEEE Access* 7:70919–70932. <https://doi.org/10.1109/ACCESS.2019.2919337>
20. S. Voutsinas, D. Karolidis, I. Voyiatzis, and M. Samarakou, Photovoltaic faults: a comparative overview of detection and identification methods, in 2021 10th International Conference on Modern Circuits and Systems Technologies, MOCASST 2021, 2021, 8–12, <https://doi.org/10.1109/MOCASST52088.2021.9493369>
21. Hu Y, Cao W, Ma J, Finney SJ, Li D (2014) Identifying PV module mismatch faults by a thermography-based temperature distribution analysis. *IEEE Trans Device Mater Reliab* 14(4):951–960. <https://doi.org/10.1109/TDMR.2014.2348195>
22. Mellit A, Tina GM, Kalogirou SA (2018) Fault detection and diagnosis methods for photovoltaic systems: a review. *Renew Sustain Energy Rev* 91(February):1–17. <https://doi.org/10.1016/j.rser.2018.03.062>
23. J. Flicker and J. Johnson, Electrical simulations of series and parallel PV arc-faults, in 2013 IEEE 39th Photovoltaic Specialists Conference (PVSC), 2013. 3165–3172, <https://doi.org/10.1109/PVSC.2013.6745127>
24. Alam MK, Khan F, Johnson J, Flicker J (2015) A comprehensive review of catastrophic faults in PV arrays: types, detection, and mitigation techniques. *IEEE J Photovoltaics* 5(3):982–997. <https://doi.org/10.1109/JPHOTOV.2015.2397599>
25. M. Sabbaghpur Arani, M. A. Hejazi, The comprehensive study of electrical faults in PV arrays, *J Electr Comput Eng*. 2016, 2016, <https://doi.org/10.1155/2016/8712960>
26. J. Johnson et al., Differentiating series and parallel photovoltaic arc-faults, in 2012 38th IEEE Photovoltaic Specialists Conference, 2012, 720–726, <https://doi.org/10.1109/PVSC.2012.6317708>
27. G. Ball et al., Inverter ground-fault detection 'blind spot' and mitigation methods. *Sol Am Board Codes Stand Rep.* no. June 2013, 2013, Available: http://solarabcs.org/about/publications/reports/blindspot/pdfs/inverter_groundfault-2013.pdf
28. J. Johnson et al., Photovoltaic DC arc fault detector testing at Sandia National Laboratories. *Conf Rec IEEE Photovolt Spec Conf*, no. June, 003614–003619, 2011, <https://doi.org/10.1109/PVSC.2011.6185930>
29. Hernández JC, Vidal PG (2009) Guidelines for protection against electric shock in PV generators. *IEEE Trans Energy Convers* 24(1):274–282. <https://doi.org/10.1109/TEC.2008.2008865>
30. W. Bower and J. Wiles, Investigation of ground-fault protection devices for photovoltaic power systems applications, *Conf. Rec. IEEE Photovolt. Spec. Conf.*, 2000-Janua, no. d, 1378–1383, 2000, <https://doi.org/10.1109/PVSC.2000.916149>
31. N. R. E. Laboratory, S. A. Sandia National Laboratory, and the S. N. L. M. Partnership, and (SuNLaMP) PV O&M Best Practices Working Group, Best practices for operation and maintenance of photovoltaic and energy storage systems ; 3rd Edition., Nrel, no. December, 153, 2018., Available: <https://www.nrel.gov/research/publications.html>
32. C. E. Packard, J. H. Wohlgenuth, and S. R. Kurtz (2012) Development of a visual inspection Checklist for evaluation of fielded PV module condition (Presentation), article, March 1, ; Golden, Colorado. <https://digital.library.unt.edu/ark:/67531/metadc829553/>
33. Köntges M et al (2017) Assessment of photovoltaic module failures in the field
34. Cubukcu M, Akanalci A (2020) Real-time inspection and determination methods of faults on photovoltaic power systems by thermal imaging in Turkey. *Renew Energy* 147:1231–1238. <https://doi.org/10.1016/j.renene.2019.09.075>
35. Madeti SR, Singh SN (2017) A comprehensive study on different types of faults and detection techniques for solar photovoltaic system. *Sol Energy* 158:161–185
36. J. Haney and A. Burstein, PV system operations and maintenance fundamentals solar, *Sol. Am. Board Codes Stand.*, no. August, 2013., Available: www.solarabcs.org
37. E. Kaplani, Detection of degradation effects in field-aged c-Si solar cells through IR thermography and digital image processing, *Int. J. Photoenergy*, 2012, 2012, <https://doi.org/10.1155/2012/396792>

38. Tina GM, Cosentino F, Ventura C (2016) Monitoring and diagnostics of photovoltaic power plants. In: Sayigh A (ed) *Renewable Energy in the Service of Mankind Vol II: Selected Topics from the World Renewable Energy Congress WREC 2014*. Springer International Publishing, Cham, pp 505–516
39. Solmetric, "Guide To Interpreting I-V Curve Measurements of PV Arrays," Appl. Note PVA-600–1, 23, 2011
40. Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H (2018) "State-of-the-art in artificial neural network applications: a survey. *Heliyon* 4(11):e00938–e00938. <https://doi.org/10.1016/j.heliyon.2018.e00938>
41. Chen Z, Wu L, Cheng S, Lin P, Wu Y, Lin W (2017) Intelligent fault diagnosis of photovoltaic arrays based on optimized kernel extreme learning machine and I-V characteristics Appl. Energy. 204:912–31. <https://doi.org/10.1016/j.apenergy.2017.05.034>. (2018)
42. Harrou F, Sun Y, Taghezouit B, Saidi A, Hamlati ME (2018) Reliable fault detection and diagnosis of photovoltaic systems based on statistical monitoring approaches. *Renew Energy* 116:22–37. <https://doi.org/10.1016/j.renene.2017.09.048>
43. Voutsinas S, Karolidis D, Voyiatzis I, Samarakou M (2022) Development of a multi-output feed-forward neural network for fault detection in Photovoltaic Systems. *Energy Rep* 8(May):33–42. <https://doi.org/10.1016/j.egyr.2022.06.107>
44. Yi Z, Etemadi AH (2017) Line-to-line fault detection for photovoltaic arrays based on multiresolution signal decomposition and two-stage support vector machine. *IEEE Trans Ind Electron* 64(11):8546–8556. <https://doi.org/10.1109/TIE.2017.2703681>
45. Xia K, He S, Tan Y, Jiang Q, Xu J, Yu W (2019) Wavelet packet and support vector machine analysis of series DC ARC fault detection in photovoltaic system. *IEEJ Trans Electr Electron Eng* 14(2):192–200. <https://doi.org/10.1002/tee.22797>
46. Harrou F, Dairi A, Taghezouit B, Sun Y (2019) An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class Support Vector Machine. *Sol Energy* 179:48–58. <https://doi.org/10.1016/j.solener.2018.12.045>. (2018)
47. Wang L, Liu J, Guo X, Yang Q, Yan W (2017) Online fault diagnosis of photovoltaic modules based on multi-class support vector machine, Proc. - 2017 Chinese Autom. Congr CAC. 2017:4569–4574. <https://doi.org/10.1109/CAC.2017.8243586>
48. Winston DP et al (2021) Solar PV's micro crack and hotspots detection technique using NN and SVM. *IEEE Access* 9:127259–127269. <https://doi.org/10.1109/ACCESS.2021.3111904>
49. Yi Z, Etemadi AH (2017) Fault detection for photovoltaic systems based on multi-resolution signal decomposition and fuzzy inference systems. *IEEE Trans Smart Grid* 8(3):1274–1283. <https://doi.org/10.1109/TSG.2016.2587244>
50. Memon SA, Javed Q, Kim WG, Mahmood Z, Khan U, Shahzad M (2022) A machine-learning-based robust classification method for PV panel faults. *Sensors* 22(21):1–14. <https://doi.org/10.3390/s22218515>
51. Jia F, Luo L, Gao S, Ye J (2019) Logistic regression based arc fault detection in photovoltaic systems under different conditions. *J Shanghai Jiaotong Univ* 24(4):459–470. <https://doi.org/10.1007/s12204-019-2095-1>
52. Fadhel S et al (2018) Data-driven approach for isolated PV shading fault diagnosis based on experimental I-V curves analysis. *Proc IEEE Int Conf Ind Technol* 2018:927–932. <https://doi.org/10.1109/ICIT.2018.8352302>
53. S. Dai, D. Wang, W. Li, Q. Zhou, G. Tian, H. Dong, Fault diagnosis of data-driven photovoltaic power generation system based on deep reinforcement learning. *Math Probl Eng* 2021 2021 <https://doi.org/10.1155/2021/2506286>
54. N. M. A. Alrahim Shannan, N. Z. Yahaya, B. Singh. Single-diode model and two-diode model of PV modules: a comparison, 2013. <https://doi.org/10.1109/ICCSC.2013.6719960>
55. De Soto W, Klein SA, Beckman WA (2006) Improvement and validation of a model for photovoltaic array performance. *Sol Energy* 80(1):78–88. <https://doi.org/10.1016/j.solener.2005.06.010>
56. Huld T, Müller R, Gambardella A (2012) A new solar radiation database for estimating PV performance in Europe and Africa. *Sol Energy* 86(6):1803–1815. <https://doi.org/10.1016/j.solener.2012.03.006>
57. Pedregosa F et al (2011) Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*. 12(85):2825–2830. Available: <http://jmlr.org/papers/v12/pedregosa11a.html>
58. Holmgren WF, Hansen CW, Mikofski MA (2018) Pvlb Python: a Python Package for Modeling Solar Energy Systems. *J Open Source Softw* 3(29):884. <https://doi.org/10.21105/joss.00884>
59. SCH, SCH6P - Multicrystalline Solar Cells.pdf. Available: <https://cdn.ensolar.com/Product/pdf/Cell/51831a51cc9b8.pdf>